National Institute of Standards and Technology
U.S. Department of Commerce

Demographic Effects in Face Recognition

Patrick Grother
Information Technology Laboratory
National Institute of Standards and Technology
United States Department of Commerce

IFPC 2018
NIST November 27, 2018

Talking about face recognition bias, quantitatively

National Institute of Standards and Technology
U.S. Department of Commerce

- Bias at what stage:
  - Capture (camera, or sub-system): **Failure-to-capture rate, Quality**
  - Template creation (for enrollment, or recognition): **Failure-to-extract rate**
  - Type II error rates: Failure to associate person with prior sample **FNMR, FNIR**
  - Type I error rates: Incorrect association of photo from two people **FMR, FPIR**

- Impact:
  - Is application dependent, so harm in one application is benefit in another
  - Magnitude matters!
    1. Demographic differentials "Δ_{AB}" matter. For example, it's bad if $|\ FNMR_A - FNMR_B\ | > \delta$
    2. Absolute error rates matter: For example, it's bad if $FNMR \gg 0$ is bad

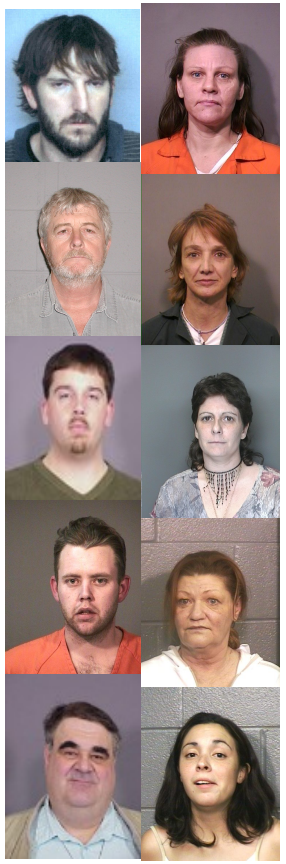
- Algorithms:
  - May differ in their biases
  - Know-Your-Algorithm
    1. Demographic differentials
    2. Other sensitivities

National Institute of Standards and Technology
U.S. Department of Commerce

1:N Face Female-Male Differential Impact

Male-Female Demographic Differential Experiment

National Institute of Standards and Technology
U.S. Department of Commerce



GALLERY:

1. N = 1,600,000
2. WHITE
3. MUGSHOTS
4. BALANCED
    1. 800,000 MALE
    2. 800,000 FEMALE
5. AGE 21-40 at FIRST ENROLLMENT

MATED PROBES:

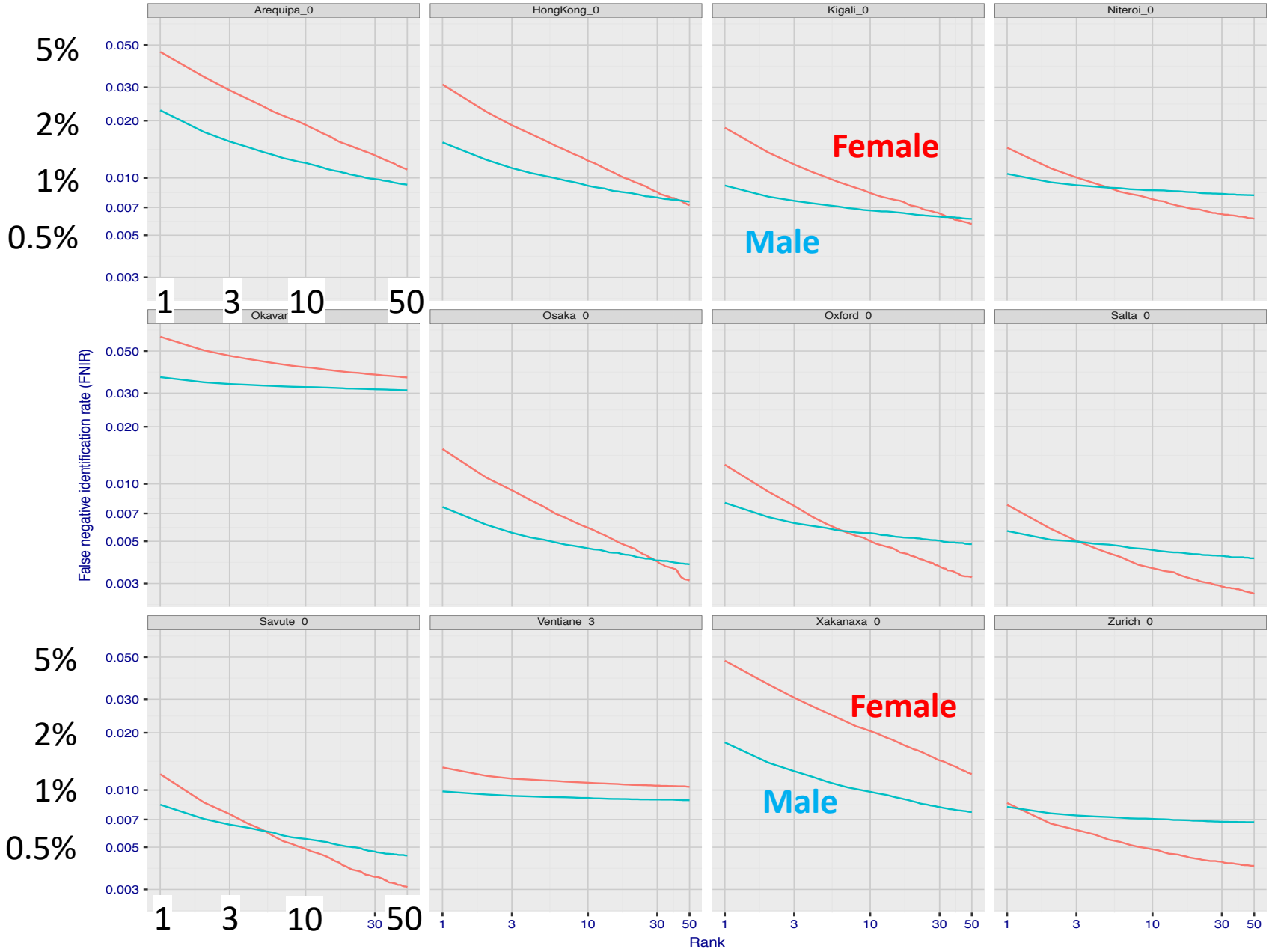
1. 100,000 MALE
2. 100,000 FEMALE
3. COLLECTED IN DIFFERENT YEAR TO GALLERY MATE
4. COLLECTED WITHIN FIVE YEARS OF MATE
5. AGE 21-40

NON-MATED PROBES:

1. 100,000 MALE
2. 100,000 FEMALE
3. AGE 21-40

Male and Female Miss Rates, FNIR(Rank)

National Institute of Standards and Technology
tment of Commerce



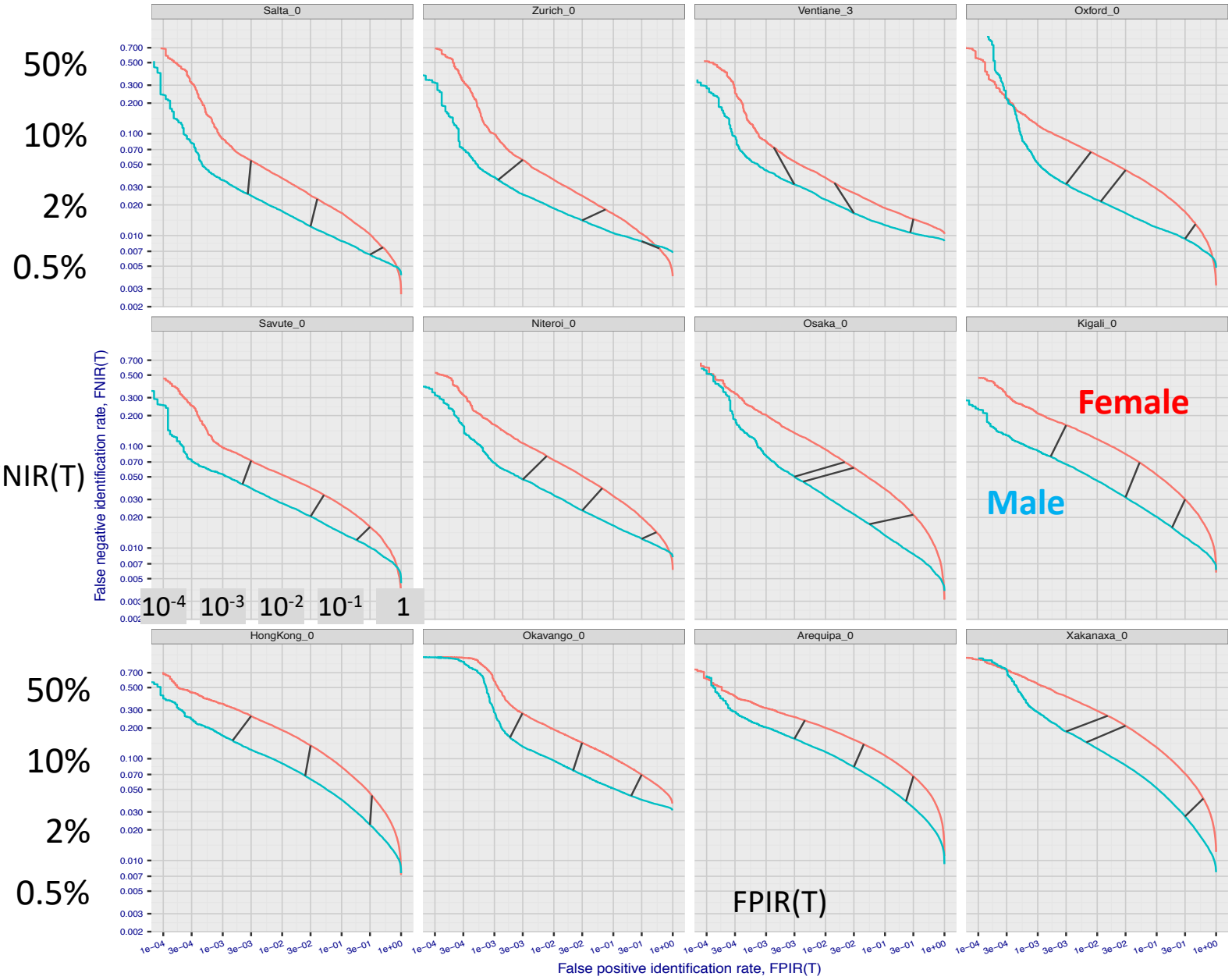
FNIR(Rank) is a metric appropriate to investigational applications where human reviewers will adjudicate candidate lists

Dataset: Mugshots MF/White/20–40/T<5yrs
Tier=1, N=1600000
F
M

N = 1 600 000 subjects,
800 000 each sex.
Enrolled with 1 image each

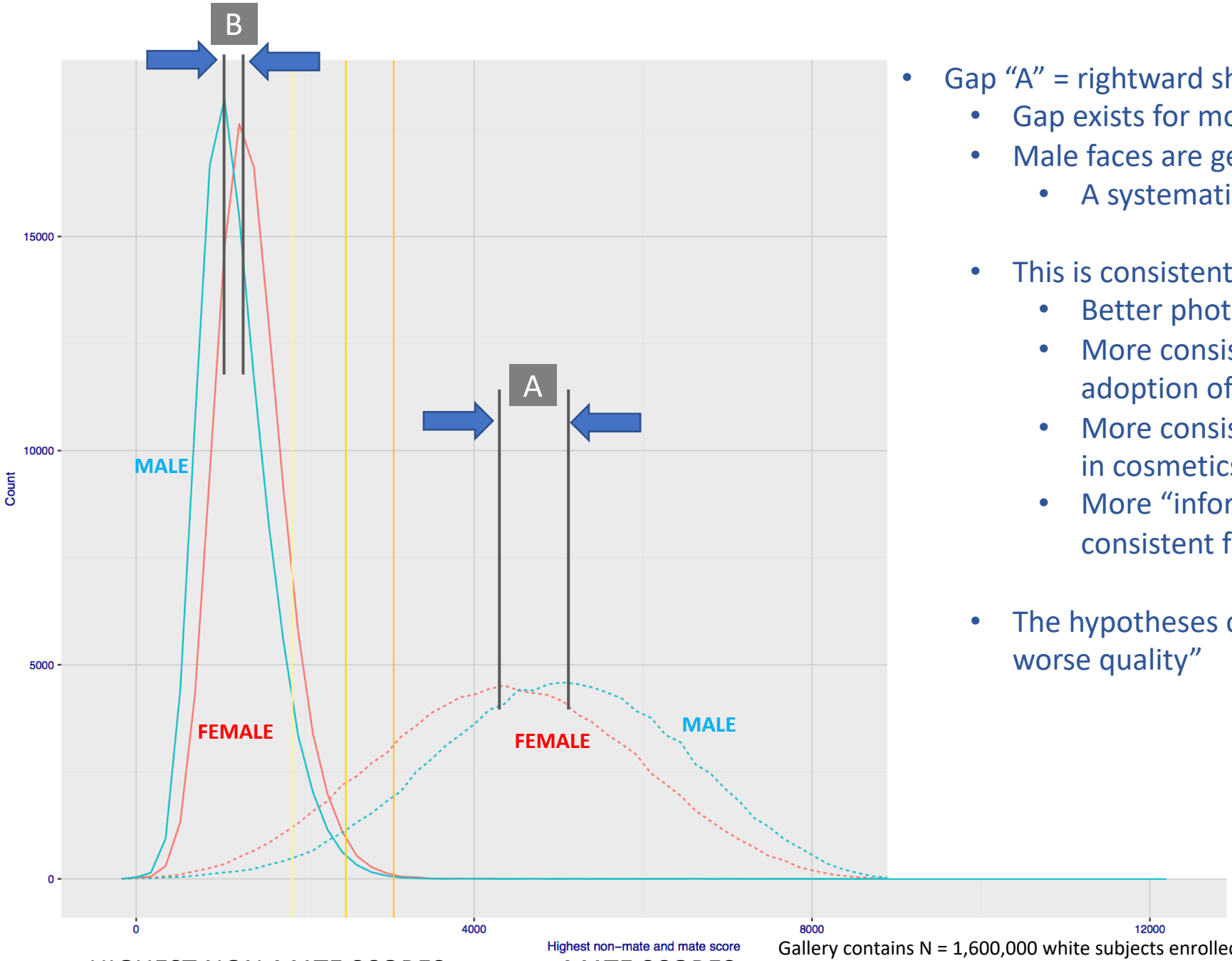
Male and Female miss rates at non-zero threshold, FNIR(T)

National Institute of Standards and Technology
U.S. Department of Commerce



FNIR(Threshold) is a metric appropriate to "identification" applications where (false) positives must be limited to available labor supply

N = 1 600 000 subjects,
800 000 each sex.
Enrolled with 1 image each

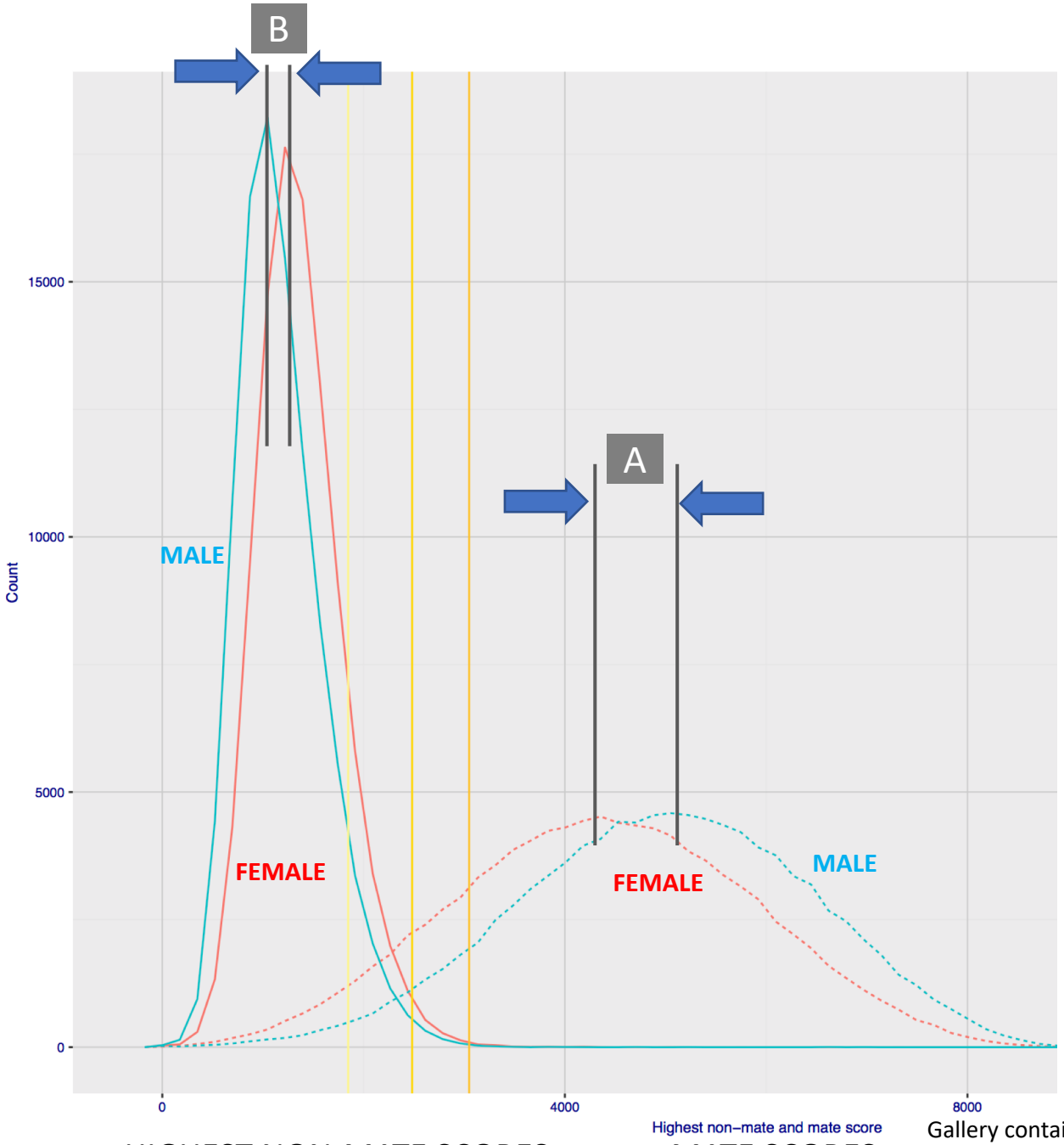


HIGHEST NON-MATE SCORES

MATE SCORES

- Gap "A" = rightward shift of male mate distribution vs. female.
  - Gap exists for most recognition algorithms
  - Male faces are generally more similar to themselves
    - A systematic effect, not confined to left tail.
  - This is consistent with, but may not actually be caused by:
    - Better photo quality in males
    - More consistent presentation to camera, e.g. adoption of frontal pose
    - More consistent condition of face, e.g. fewer changes in cosmetics or eyewear
    - More "information-rich" faces, e.g. presence of consistent features.
  - The hypotheses can be inverted e.g. "females' photos have worse quality"

Gallery contains N = 1,600,000 white subjects enrolled with a single image, 800,000 each of men and women. Age, at enrollment is 21-40. The probe images are collected later, in a different calendar year, and within five years.



HIGHEST NON-MATE SCORES

MATE SCORES

- Gap "B" = rightward displacement of female non-mate distribution relative to male.
  - This gap occurs for most recognition algorithms
  - The shift indicates that when a female face is searched against a gallery of different people, balanced 50-50 male and female, it tends to yield higher scores.
  - Displacement of the entire distribution shows a systematic effect, not confined to the right tail of the distribution.

  - This is consistent with, but may not actually be caused by:
    - Nature: Female faces are naturally more similar to each other
    - Photographic effects: Females' photos include some artifacts that algorithms match but should not e.g. hairstyles.

- If errors are all that matters, are the score distribution displacements important? They reveal anatomic or photographic interactions with the algorithms.

- This experiment uses only images of whites; different effects might occur in other races

Gallery contains N = 1,600,000 white subjects enrolled with a single image, 800,000 each of men and women. Age, at enrollment is 21-40. The probe images are collected later, in a different calendar year, and within five years.

National Institute of Standards and Technology
U.S. Department of Commerce

1:N Face Black-White Differential Impact

Black-white Demographic Difference Experiment

National Institute of Standards and Technology
U.S. Department of Commerce

GALLERY:

1. N = 1,600,000
2. MALE
3. MUGSHOTS
4. BALANCED
    1. 800,000 BLACK
    2. 800,000 WHITE
5. AGE 21-40 at FIRST ENROLLMENT

MATED PROBES:

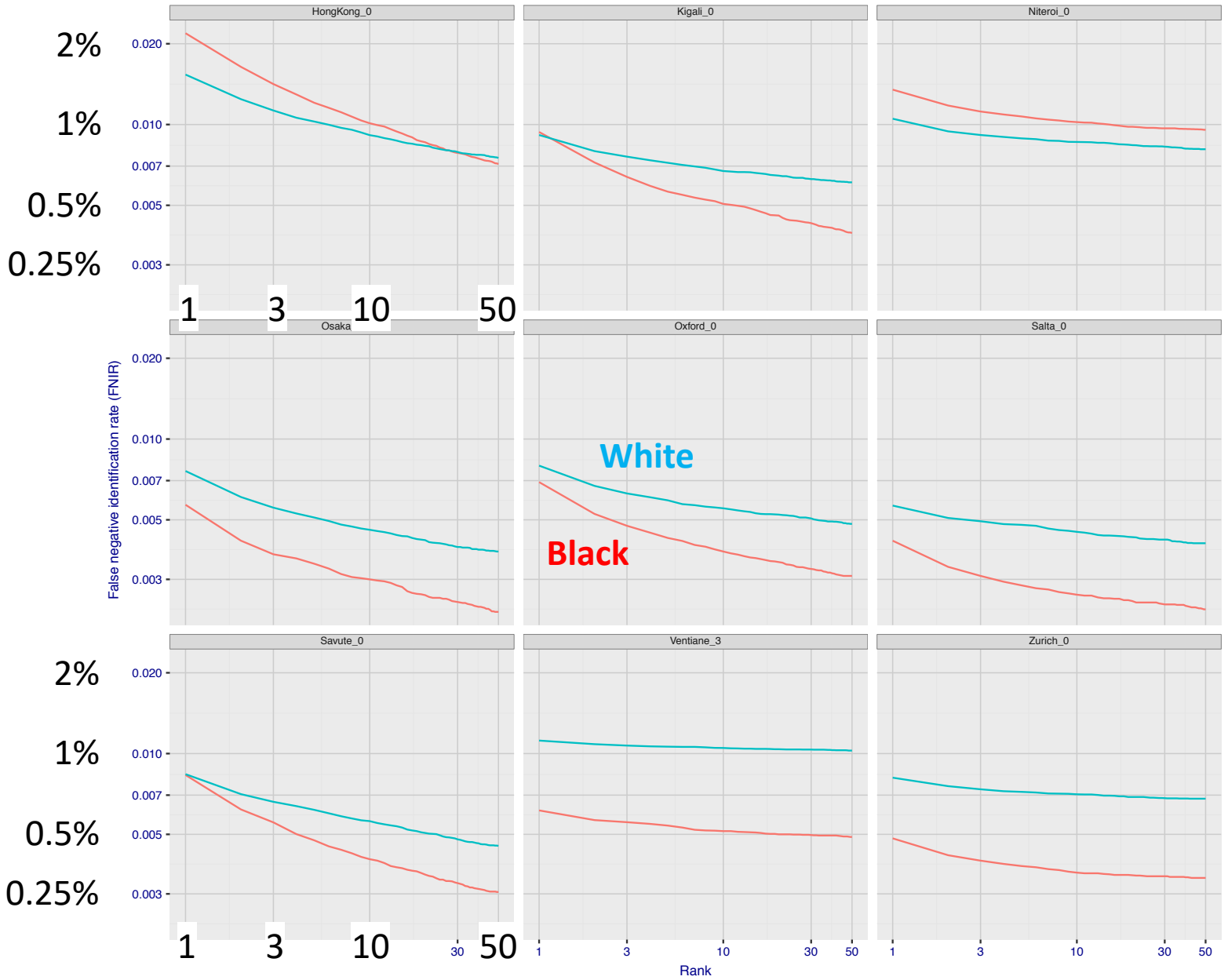
1. 100,000 BLACK
2. 100,000 WHITE
3. COLLECTED IN DIFFERENT YEAR TO GALLERY MATE
4. COLLECTED WITHIN FIVE YEARS OF MATE
5. AGE 21-40

NON-MATED PROBES:

1. 100,000 BLACK
2. 100,000 WHITE
3. AGE 21-40

Black and White Miss Rates: FNIR(Rank)

National Institute of Standards and Technology
U.S. Department of Commerce

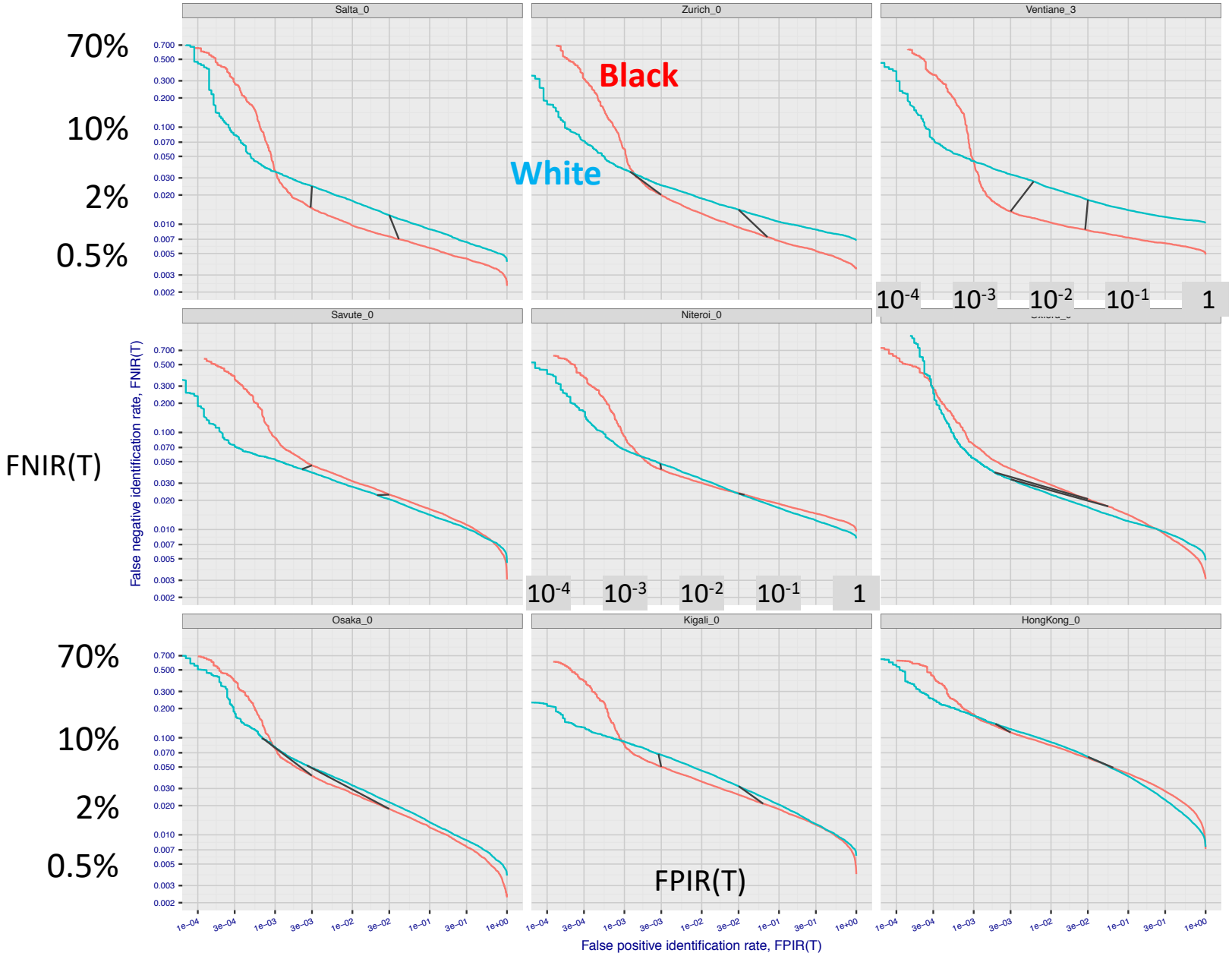


FNIR(Rank) is a metric appropriate to investigational applications where human reviewers will adjudicate candidate lists

N = 1 600 000 subjects, 800 000 each race. Enrolled with 1 image each

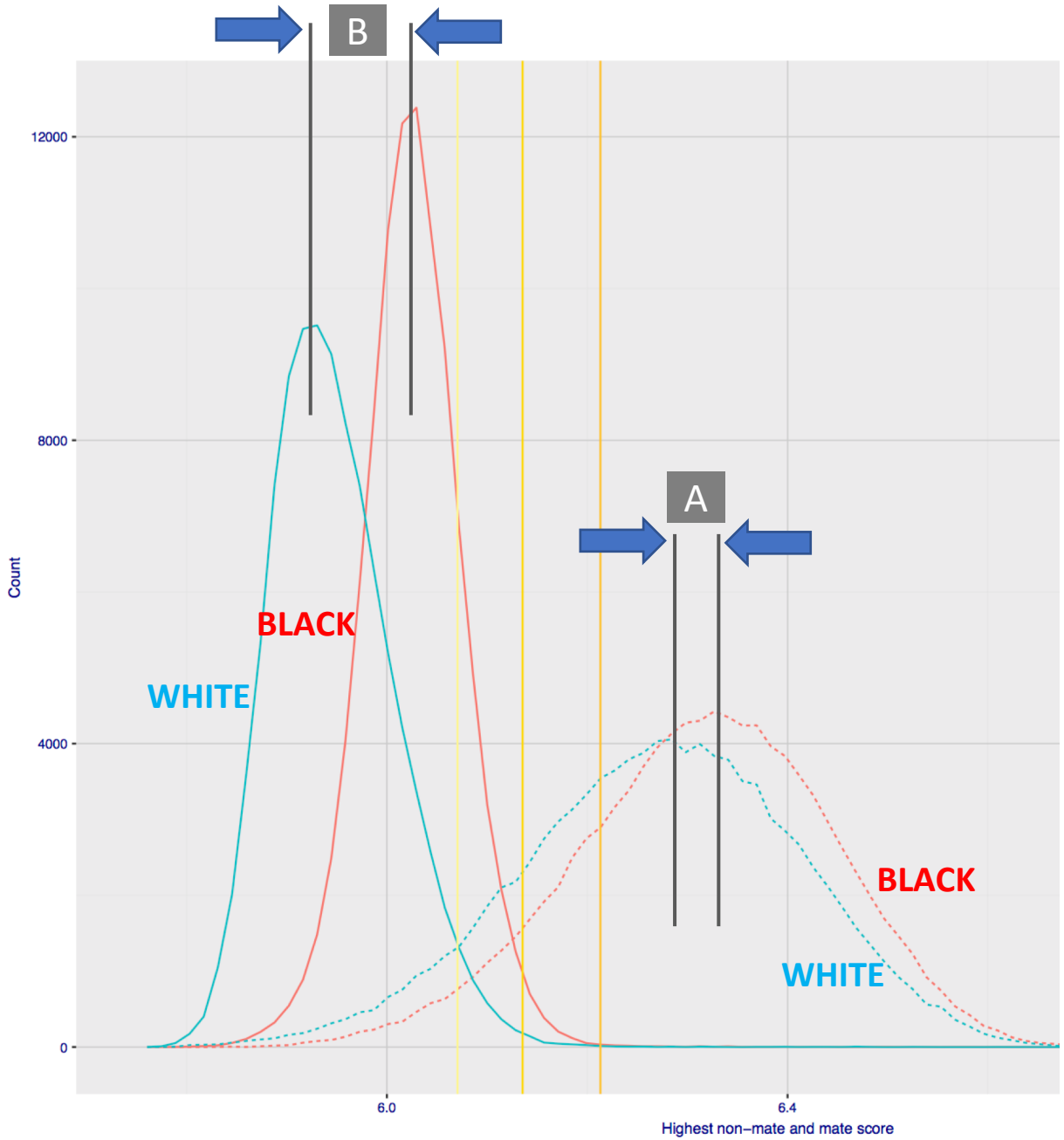
Black White Miss Rates at Non-Zero Threshold, FNIR(T)

National Institute of Standards and Technology
U.S. Department of Commerce



FNIR(Threshold) is a metric appropriate to "identification" applications where (false) positives must be limited to available labor supply

N = 1 600 000 subjects, 800 000 each race. Enrolled with 1 image each



HIGHEST NON-MATE SCORES

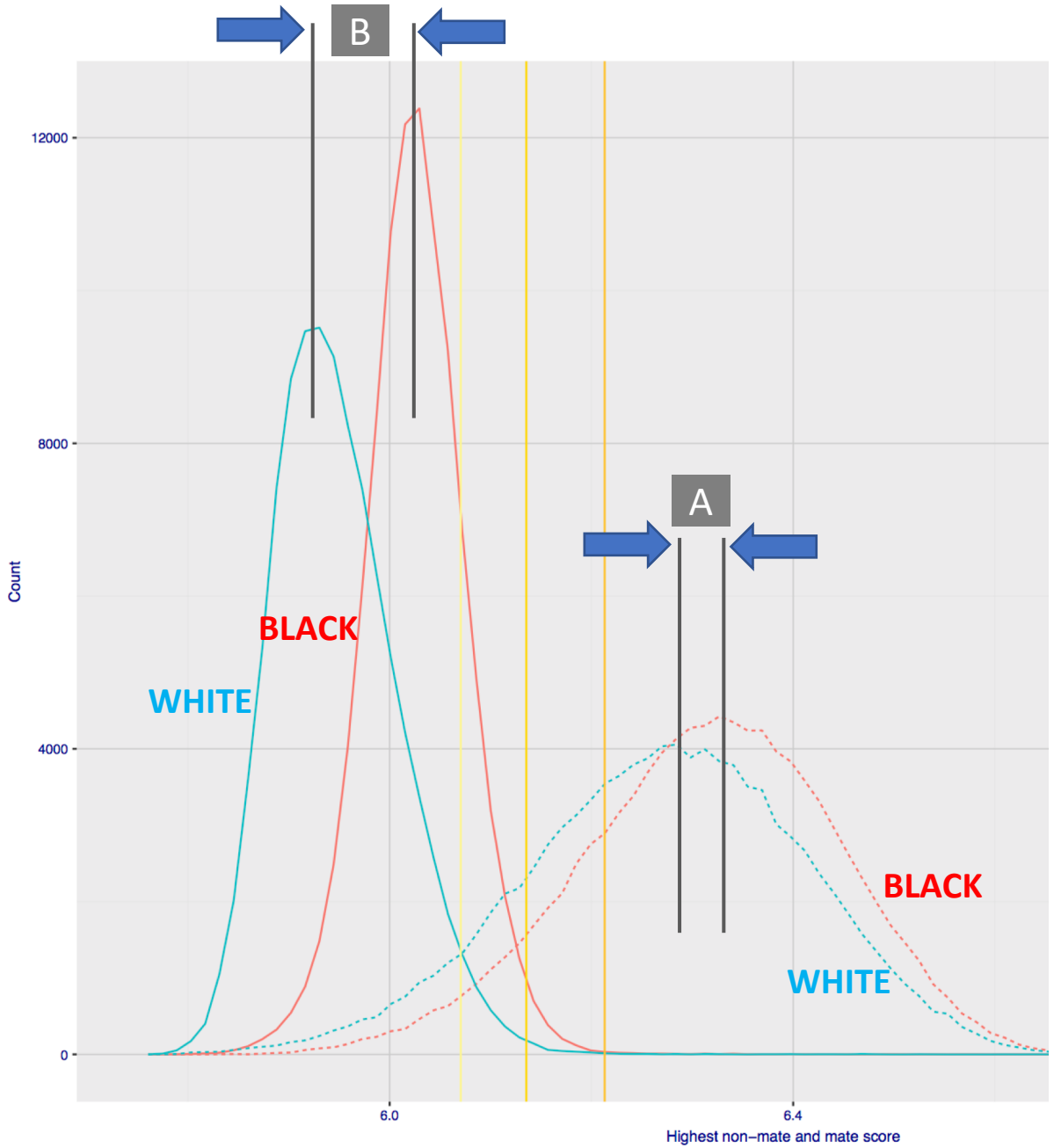
MATE SCORES

- Gap"A" = rightward displacement of black mate distribution relative to white.
  - This gap occurs for many recognition algorithms, including some of the more accurate algorithms
  - Black faces are generally more similar.
  - Displacement of distribution shows a systematic effect, not confined only to the left tail.

- This is consistent with, but may not actually be caused by:
  - Better photo quality
  - More consistent presentation to camera, e.g. adoption of frontal pose
  - More consistent condition of face, e.g. fewer changes in beard presence
  - More "information-rich" faces, e.g. presence of consistent features.

- The hypotheses can be inverted e.g. "white photos have worse quality"

Gallery contains N = 1,600,000 males enrolled with a single image, 800,000 each of black and white. Age, at enrollment is 21-40. The probe images are collected later, in a different calendar year, and within five years of enrollment.



HIGHEST NON-MATE SCORES

MATE SCORES

- Gap "B" = rightward shift of black non-mate distribution vs white.
  - This gap occurs for most recognition algorithms
  - The shift indicates that when a black face is searched against a gallery of different people, balanced 50-50 black and white, it tends to yield higher scores.
  - The displacement of the entire distribution shows a systematic effect, it is not confined only to the right tail of the distribution.
  - This is consistent with, but may not actually be caused by:
    - Nature: Photos of black subjects' faces are naturally more similar to each other
    - Photographic effects: Photos of black subjects' faces include some artifacts that match each other e.g. specular reflections.
    - ...
- If errors are all that matters, are the score distribution displacements important? They reveal anatomic or photographic interactions with the algorithms.
- This experiment uses only images of men. Different effects may occur for women.

Gallery contains N = 1,600,000 males enrolled with a single image, 800,000 each of black and white. Age, at enrollment is 21-40. The probe images are collected later, in a different calendar year, and within five years of enrollment.

Next steps

National Institute of Standards and Technology
U.S. Department of Commerce

- NIST has given some quantitative feedback to developers
  - 2018-05 and 2018-09
  - Ongoing and expanded tests

- NIST will publish an Interagency Report Q1 2019 on demographics effects in face recognition

  - Existing content for 1:1 algorithms in FRVT Ongoing Reports
  - https://www.nist.gov/programs-projects/face-recognition-vendor-test-frvt-ongoing

Thanks

patrick.grother@nist.gov